Sampling from a Finite Population: Interval Estimation of Means, Proportions and Population Totals

Jerry Brunner

March 21, 2007

Most of the material in this course is based on the assumption that we are sampling with replacement, or else sampling without replacement from an "infinite population" (definitely a theoretical abstraction. We have justified this on the grounds of simplicity, and also because if the population is very large, there is very little difference between sampling with and without replacement.

But in practice, sampling is almost always without replacement. Furthermore, we can use information about the size of the population (and sometimes the sizes of sub-populations) to estimate population totals, and to make the estimation of means and percentages more precise. Precision of estimation is usually purchased by increasing sample size, and data collection costs money. If you can design a survey so as to get precise estimation some other way, you can use a smaller sample and save money.

1 Simple random sampling without replacement

Suppose we select a random sample of size n without replacement from a population of size N. Imagine putting a N balls in a jar; each ball is labelled with the identification number of one member of the population. You pull out n balls without looking (and without replacement), record the numbers, and collect data from the corresponding members of the population. the population mean is μ, and the population standard deviation is σ.

Each subset of the population will have an equal chance of being chosen, and each individual in the population will have an equal chance of being in the sample. Of course in practice you would not use a big jar of balls; you'd model the process of selection on a computer, using a stream of pseudo-random numbers from a random number generator.

1.1 Estimating means

When you randomly sample n units without replacement from a population of size N,

- The point estimate of the population mean μ is still the sample mean $\overline{x}$.

- $\overline{x}$ is still unbiased for μ. That is, the expected value of the sampling distribution of $\overline{x}$ is μ.

- The standard error of the mean (standard deviation of the sampling distribution of $\overline{x}$) is $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$. We never know σ, but for $n \geq 30$ we can estimate it with the sample standard deviation s, and estimate $\sigma_{\overline{x}}$ with
$$\widehat{\sigma}_{\overline{x}} = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \tag{1}$$

- The Central Limit Theorem still applies. We will assume $n \geq 30$, with a different rule when the population mean is actually a proportion.

Notice that when you are sampling *with* replacement, the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$, but when you are sampling *without* replacement, the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ multiplied by $\sqrt{1-\frac{n}{N}}$. The quantity $\sqrt{1-\frac{n}{N}}$ is called the **finite population correction factor**.

- The finite population correction factor is always less than one. This means that when you sample without replacement, estimation of the population mean is a little more precise; the margin of error is a bit smaller, and the confidence interval is a bit narrower.

- If $n = N$, the finite population correction factor equals zero, and so does $\sigma_{\overline{x}}$. This makes sense. If you sample the whole population, there is no error of estimation.

Consider a population with true standard deviation $\sigma = 6$. We estimate μ with the sample mean $\overline{x}$, based on a random sample of size n without replacement from a population of size N. Table 1.1 shows the standard error of the mean $\sigma_{\overline{x}}$ for various values of n and N. Bear in mind that $\sigma_{\overline{x}}$ represents the variation of $\overline{x}$ around the population mean, so smaller values of $\sigma_{\overline{x}}$ mean more precise estimation.

Notice that once the population is reasonably large, $\sigma_{\overline{x}}$ barely changes as the population size gets larger, while it goes down rapidly with increasing sample size. The moral is that except for very small populations, *what matters is the size of the sample, not the size of the population.*

The Central Limit Theorem for finite populations says that for large samples,
$$Z = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}$$
is approximately standard normal; we will apply it when $n \geq 30$. This leads to the estimated $(1-\alpha)100\%$ margin of error
$$z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \tag{2}$$
and the $(1-\alpha)100\%$ confidence interval for μ
$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \tag{3}$$

Table 1: Standard error of the mean $\sigma_{\overline{x}}$, sampling n observations without replacement from a population of size N: Population standard deviation is $\sigma = 20$

	n				
N	25	50	100	500	1,000
50	2.8284	0.0000			
100	3.4641	2.0000	0.0000		
500	3.8987	2.6833	1.7889	0.0000	
1,000	3.9497	2.7568	1.8974	0.6325	0.0000
10,000	3.9950	2.8213	1.9900	0.8718	0.6000
100,000	3.9995	2.8277	1.9990	0.8922	0.6293
1,000,000,000	4.0000	2.8284	2.0000	0.8944	0.6325
10,000,000,000	4.0000	2.8284	2.0000	0.8944	0.6325
100,000,000,000	4.0000	2.8284	2.0000	0.8944	0.6325

1.2 Estimating proportions and percentages

Suppose the data $x_1, \ldots, x_n$ are coded so that $x_i = 1$ if the event (the customer buys the product, the shipment arrives on time, etc.) happens for observation i, and $x_i = 0$ if it does not happen. Then

- The population mean is $\mu = p$, the population proportion.
- The sample mean is $\overline{x} = \widehat{p}$, the sample proportion
- The population standard deviation is $\sigma = \sqrt{p(1-p)}$, but we never know it, so we use the estimate $\sqrt{\widehat{p}(1-\widehat{p})}$.
- The estimated standard deviation of the sample proportion is
$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{1-\frac{n}{N}}. \tag{4}$$
- The Central Limit Theorem applies directly, because we are really dealing with a sample mean. For large samples, $Z = (\widehat{p} - p)/\widehat{\sigma}_{\widehat{p}}$ is approximately standard normal.
- Because we have binary data, we don't use the $n \geq 30$ rule. Instead, we check whether $\widehat{p} \pm 3\widehat{\sigma}_{\widehat{p}}$ is inside the interval from zero to one.

The estimated $(1-\alpha)100\%$ margin of error for $\widehat{p}$ as an estimate of p is

$$z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{1-\frac{n}{N}} \tag{5}$$

and the $(1-\alpha)100\%$ confidence interval for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{1-\frac{n}{N}}. \tag{6}$$

To obtain a margin of error or confidence interval for a percentage, just multiply everything by 100. That is, the estimated $(1-\alpha)100\%$ margin of error for the sample percentage as an estimate of the population percentage is

$$100 \times z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{1-\frac{n}{N}} \tag{7}$$

and the $(1-\alpha)100\%$ confidence interval for p is

$$100 \times \widehat{p} \pm 100 \times z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\sqrt{1-\frac{n}{N}}. \tag{8}$$

1.3 Estimating population totals

Since

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

the population total $\sum_{i=1}^{N} x_i = N\mu$, and we estimate it with $N\overline{x}$. This applies whether data are binary or not. If the data are zeros and ones, the population total is the total number of something, like the total number of cable TV subscribers in Canada. If the data represent amount of something, the population total is total amount, like the total amount of money spent on fast food in Ontario during the past month. If the data are something like ratings of customer satisfaction, the population mean is of interest, but the population total is meaningless.

Using the symbol $\overline{x}$ instead of $\widehat{p}$ if the data happen to be zeros and ones, the $(1-\alpha)100\%$ margin of error for the estimated population total is

$$N z_{\alpha/2}\widehat{\sigma}_{\overline{x}} \tag{9}$$

and the $(1-\alpha)100\%$ confidence interval for the population total is

$$N\overline{x} \pm N z_{\alpha/2}\widehat{\sigma}_{\overline{x}}, \tag{10}$$

where again, if the data are 1=Yes and 0=No, then $\overline{x} = \widehat{p}$, and $\widehat{\sigma}_{\overline{x}}$ is given by (16). Otherwise, $\widehat{\sigma}_{\overline{x}}$ is given by (1).

Example 1.1 *The United States is divided into 3078 districts (counties or county equivalents). We randomly sample 300 of these without replacement, and determine the number of acres devoted to farms. The sample mean number of farm acres is 297.9 thousand, with a standard deviation of 344.6 thousand. Give a point estimate and a 95% margin of error for the total number of acres devoted to farms in the U. S.*

Exercises

1. We select a random sample of 300 without replacement from a population of size 450. What is the finite population correction factor? The answer is a number.

2. From a campus with 4250 students, we select a random sample of size 200 without replacement, and ask if they have broadband Internet access at home; 145 say Yes.

   (a) Give a 95% confidence interval for the population proportion of students who claim to have Internet access at home.

   (b) Give a 95% confidence interval for the true *percentage* of students who claim to have Internet access at home.

   (c) Give a point estimate and a 95% confidence interval for the true *total number* of students who claim to have Internet access at home.

   (d) How do you know that the sample size is large enough to do what you have done? Show your work.

3. There are 31,989 homes in Stephens County, a rural area in Manitoba. We select a simple random sample of 500 homes without replacement, and check their assessed value from county records. We get a sample mean of \$69,368 and a standard deviation of \$16,456.

   (a) Give a 95% confidence interval for the population mean assessed value of homes in Stephens County.

   (b) Give a point estimate and a 95% confidence interval for the true *total* assessed value of homes in Stephens County.

   (c) How do you know that the sample size is large enough to do what you have done?

2 Stratified random sampling

In stratified random sampling, the population is divided into segments, or *strata*. The sizes of the strata are known, usually from census data. Then you select a probability sample (in our case, a simple random sample without replacement) from each stratum, with the objective of estimating population means, percentages and totals.

Why would you stratify? If estimates within strata are of interest (like estimating the mean time playing video games in each province), stratification can ensure that you have enough data from each stratum to do a reasonable analysis. To accomplish this, it is common to deliberately *over-sample* from smaller strata. Another advantage of stratification is that it can increase precision when you are estimating parameters of the whole population – provided the variable you are interested in is substantially different from stratum to stratum. For example, if you were interested in estimating the average hours of sports watched by students on a campus, it would be natural to stratify by sex.

Notation We divide the population into k strata. If we are stratifying by sex, $k = 2$. If we are stratifying the Canadian population by province or territory, $k = 13$. We obtain a sample from each stratum. Here is a summary of the notation, assuming $k = 4$ strata. Of course you would not use the notation for proportions and means in the same problem.

	Stratum (j)			
	1	2	3	4
Population Size	N_1	N_2	N_3	N_4
Population Mean	μ_1	μ_2	μ_3	μ_4
Population Variance	σ_1^2	σ_2^2	σ_3^2	σ_4^2
Population Proportion	p_1	p_2	p_3	p_4
Sample Size	n_1	n_2	n_3	n_4
Sample Mean	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
Sample Variance	s_1^2	s_2^2	s_3^2	s_4^2
Sample Proportion	$\widehat{p}_1$	$\widehat{p}_2$	$\widehat{p}_3$	$\widehat{p}_4$

We will often be summing over the index j, with j running from one to k. For example, the total population size is $N = \sum_{i=1}^k N_j$, and the total sample size is $n = \sum_{i=1}^k n_j$.

2.1 Estimating means

Using some algebra that will be skipped here, it can be shown that the overall population mean is a sort of weighted average of the stratum population means.

$$\mu = \sum_{j=1}^k \left(\frac{N_j}{N}\right) \mu_j = \left(\frac{N_1}{N}\right) \mu_1 + \left(\frac{N_2}{N}\right) \mu_2 + \cdots + \left(\frac{N_k}{N}\right) \mu_k$$

The population mean is estimated by the corresponding weighted average of sample means.

$$\widehat{\mu} = \sum_{j=1}^k \left(\frac{N_j}{N}\right) \overline{x}_j \qquad (11)$$

It is natural that the sub-population sample means would be weighted so as to give more importance to the ones from the larger strata. In any case, the estimator $\widehat{\mu}$ is unbiased, and if the sample sizes of the strata are all above 25, the Central Limit Theorem applies and it has a sampling distribution that is approximately normal.

The standard error of $\widehat{\mu}$ (that is, the estimated standard deviation of its sampling distribution) is given by

$$\widehat{\sigma}_{\widehat{\mu}} = \sqrt{\sum_{j=1}^k \left(\frac{N_j}{N}\right)^2 \frac{s_j^2}{n_j} \left(1 - \frac{n_j}{N_j}\right)} \qquad (12)$$

Here's how you calculate this formula. For each stratum, multiply three terms together. The first term is $\left(\frac{N_j}{N}\right)^2$, the proportion of the population that is in stratum j, but squared.

The second term is $\frac{s_j^2}{n_j}$, the standard error of the sample mean for stratum j, but squared. The third term is $\left(1 - \frac{n_j}{N_j}\right)$, the finite population correction factor, but squared.

After taking the product of these three terms for each stratum, you add up the products. *Then* you take the square root of the sum.

Once you have the standard error, it's easy to calculate the

$$(1-\alpha)100\% \text{ Margin of Error} = z_{\alpha/2}\widehat{\sigma}_{\widehat{\mu}}, \tag{13}$$

and the confidence interval is $\widehat{\mu}$ plus or minus the margin of error. That is,

$$\widehat{\mu} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\mu}}. \tag{14}$$

Here is an example with two strata. On a quiz or the exam, you might be asked for a point estimate of μ if there are more than two strata, but a margin of error would require too much time. To calculate a margin of error for a problem with more than two strata, you need Excel or something like that; expect it as a computer assignment.

Example 2.1 *On a campus of 4,217 students, 2,241 are females and 1,976 are males. We randomly select 50 males without replacement and 50 females without replacement, and take a variety of physical measurements. For the females, we obtain a mean height of 167.6 cm, with a standard deviation of 5.1. For the males, we obtain a mean height of 177.8 cm, with a standard deviation of 7.6. Give a point estimate and a the mean height of all the students on the campus.*

It helps to write down the numbers in tabular format. The first four rows are given by the problem; the rest are calculated.

	Females	Males	Total
N_j	2,241	1,976	4,217
n_j	50	50	
$\overline{x}_j$	167.6	177.8	
s_j	5.1	7.4	
s_j^2	26.01	54.76	
$\frac{N_j}{N}$	0.53	0.47	
$\left(\frac{N_j}{N}\right)^2$	0.2809	0.2209	
$\frac{s_j^2}{n_j}\left(1 - \frac{n_j}{N_j}\right)$	0.51	1.07	
$\left(\frac{N_j}{N}\right)^2 \frac{s_j^2}{n_j}\left(1 - \frac{n_j}{N_j}\right)$	0.1433	0.2364	0.3797

Applying (11), we get a point estimate of

$$\widehat{\mu} = \left(\frac{2,241}{4,217}\right)167.6 + \left(\frac{1,976}{4,217}\right)177.8 = 172.38.$$

The last row has the products we need for the estimated standard error of $\widehat{\mu}$, and the lower right-hand cell has the sum of products. All we need to do now is take the square root, and

$$\widehat{\sigma}_{\widehat{\mu}} = \sqrt{0.3797} = 0.616.$$

The 95% margin of error is now $z_{\alpha/2}\widehat{\sigma}_{\widehat{\mu}} = (1.96)(0.616) = 1.21$, and the 95% confidence interval is $\widehat{\mu} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\mu}} = 172.38 \pm 1.21$, or the interval from 171.17cm to 173.59cm.

2.2 Estimating proportions and percentages

The formulas we need are closely parallel to the ones from the preceding section. The point estimate of a population proportion is

$$\widehat{p} = \sum_{j=1}^{k} \left(\frac{N_j}{N}\right) \widehat{p}_j, \tag{15}$$

with standard error

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\sum_{j=1}^{k} \left(\frac{N_j}{N}\right)^2 \frac{\widehat{p}_j(1-\widehat{p}_j)}{n_j} \left(1 - \frac{n_j}{N_j}\right)} \tag{16}$$

To use the normal distribution to get a margin of error, we need $\widehat{p} \pm 3\widehat{\sigma}_{\widehat{p}}$ to be inside the interval from zero to one. To estimate percentages, multiply the following by 100: the point estimate, the margin of error, and the endpoints of the confidence interval.

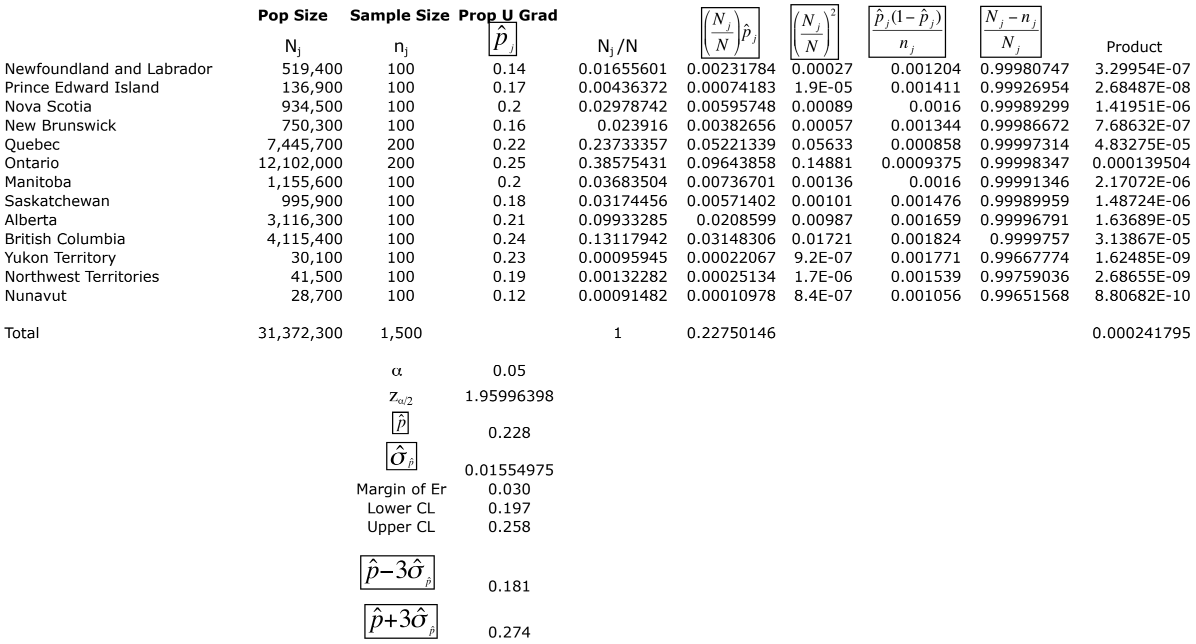
Example 2.2 *We take independent random samples without replacement from each of the Canadian Provinces and Territories. Among other things, we ask education. Figure 1 below shows the calculations needed to estimate the proportion of Canadians who finished university. This is a printout of the* Excel *spreadsheet* ***`ugrad.xls`***, *which is available online.*

Of some interest is $z_{\alpha/2}$; it definitely did not come from a table. The code is `=NORMINV(1-D19/2,0,1)`. You give the `NORMINV` function a number (call it a), and it gives you the value x such that the area under the normal curve less than or equal to x is equal to a. The second and third arguments of the function are the mean and standard deviation of the normal distribution you want. I asked for a standard normal distribution by specifying 0 and 1. The first argument is $1-\alpha/2$, where $\alpha = 0.05$ is in `D19`. If I change the value in `D19`, say to 0.01, the margin of error and the confidence interval update instantly. It's very nice.

2.3 Estimating population totals

As in the case of a single random sample, population totals are estimated by multiplying a sample proportion or sample mean by the population size. If the data are zeros and

Figure 1: Spreadsheet for Estimating Proportion of Canadians with University Educations



	Pop Size	Sample Size	Prop U Grad						
	N_j	n_j	$\boxed{\hat{p}_j}$	N_j/N	$\boxed{\left(\frac{N_j}{N}\right)\hat{p}_j}$	$\boxed{\left(\frac{N_j}{N}\right)^2}$	$\boxed{\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}}$	$\boxed{\frac{N_j-n_j}{N_j}}$	Product
Newfoundland and Labrador	519,400	100	0.14	0.01655601	0.00231784	0.00027	0.001204	0.99980747	3.29954E-07
Prince Edward Island	136,900	100	0.17	0.00436372	0.00074183	1.9E-05	0.001411	0.99926954	2.68487E-08
Nova Scotia	934,500	100	0.2	0.02978742	0.00595748	0.00089	0.0016	0.99989299	1.41951E-06
New Brunswick	750,300	100	0.16	0.023916	0.00382656	0.00057	0.001344	0.99986672	7.68632E-07
Quebec	7,445,700	200	0.22	0.23733357	0.05221339	0.05633	0.000858	0.99997314	4.83275E-05
Ontario	12,102,000	200	0.25	0.38575431	0.09643858	0.14881	0.0009375	0.99998347	0.000139504
Manitoba	1,155,600	100	0.2	0.03683504	0.00736701	0.00136	0.0016	0.99991346	2.17072E-06
Saskatchewan	995,900	100	0.18	0.03174456	0.00571402	0.00101	0.001476	0.99989959	1.48724E-06
Alberta	3,116,300	100	0.21	0.09933285	0.0208599	0.00987	0.001659	0.99996791	1.63689E-05
British Columbia	4,115,400	100	0.24	0.13117942	0.03148306	0.01721	0.001824	0.9999757	3.13867E-05
Yukon Territory	30,100	100	0.23	0.00095945	0.00022067	9.2E-07	0.001771	0.99667774	1.62485E-09
Northwest Territories	41,500	100	0.19	0.00132282	0.00025134	1.7E-06	0.001539	0.99759036	2.68655E-09
Nunavut	28,700	100	0.12	0.00091482	0.00010978	8.4E-07	0.001056	0.99651568	8.80682E-10
Total	31,372,300	1,500		1	0.22750146				0.000241795

α	0.05
$z_{\alpha/2}$	1.95996398
$\boxed{\hat{p}}$	0.228
$\boxed{\hat{\sigma}_{\hat{p}}}$	0.01554975
Margin of Er	0.030
Lower CL	0.197
Upper CL	0.258
$\boxed{\hat{p}-3\hat{\sigma}_{\hat{p}}}$	0.181
$\boxed{\hat{p}+3\hat{\sigma}_{\hat{p}}}$	0.274

ones, the population total is the total number of something. For example, one might divide the city of Toronto into little square regions, take a random sample of regions, and estimate the mean number of under-the-table "cash only" plumbers in each one. This estimate would be multiplied by the total number of regions to estimate the total number of unregistered plumbers in Toronto.

If the data consist of *amount* of something, the population total is total amount. For example, suppose we could estimate the mean amount of gasoline in the gas tanks of vehicles registered in Ontario, and then multiply by the total number of vehicles to estimate the total litres of gas in this unpublicized "strategic reserve."

To estimate population totals, start with either estimation of a proportion or a more general mean, and then multiply the following by 100: the point estimate, the margin of error, and the endpoints of the confidence interval. There; we did it without any formulas.

Example 2.3 *Continuing Example 2.2, give a point estimate and a 95% margin of error for the total number of university graduates in Canada.*

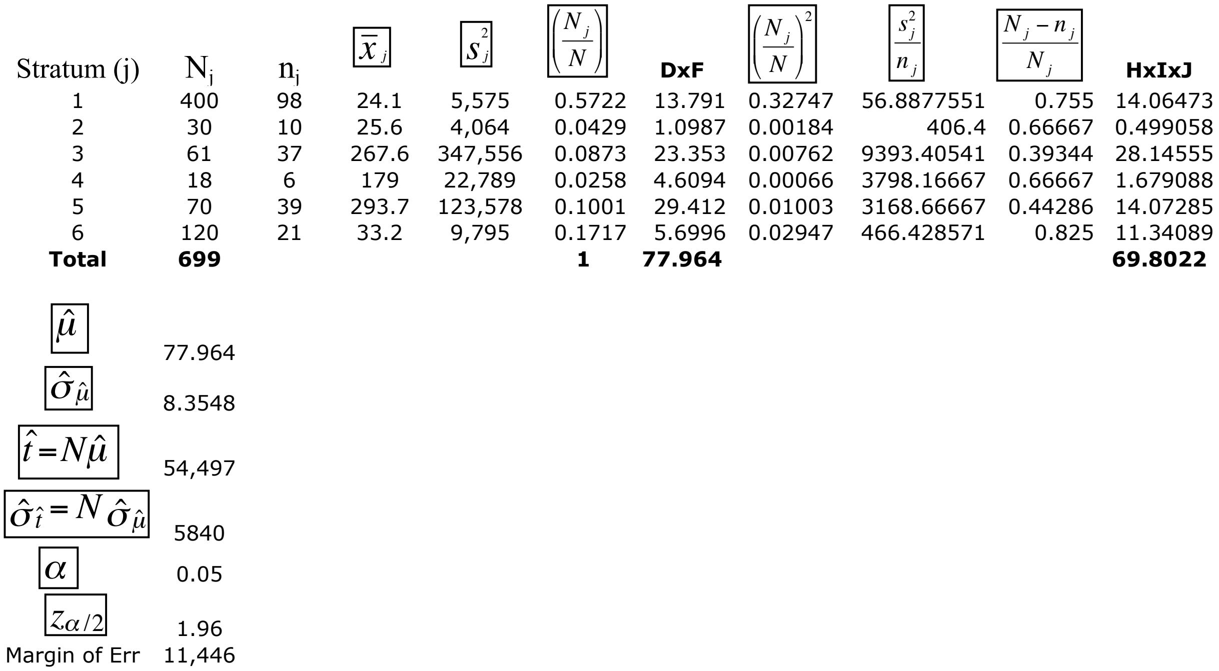
The point estimate is $\widehat{t} = N\widehat{p} = (31,372,300)(0.228) = 7,152,884$ university graduates. The margin of error is N times the 95% margin of error for the proportion. That's $(31,372,300)(0.030) = 941,169$.

That was easy, but the moral of the story is serious. That margin of error is *large!* We are comfortable with our estimate give or take almost a million people. It's not very impressive, but that's how it goes with the estimation of totals. You multiply a small margin of error by a very big number, and you get a big number. The estimation of totals is inherently not very precise. This is all the more reason to provide a margin of error along with any estimate.

Example 2.4 *Managers of a cable TV company want to know the number of satellite dishes in a large city. They have high-resolution photographs of the city from the air. They divide the city into square regions of equal area. With enough patience, it is possible to count the number of dishes in one of the squares using a magnifying glass, but it is too expensive to hire people to do it all, or to divert enough current employees from their duties to d the job. Accordingly, management decide to estimate the quantity they want.*

The city is divided into six regions (strata) based on average building height, yard size, residential/commercial mix, and so on. Then a random sample of squares is selected from each region, and the number of dishes in each square is counted from aerial photographs. The data are in the spreadsheet `TotalDishes.xls` *(See Figure 2; it's also posted online.) Give a point estimate and a 95% margin of error for the total number of satellite dishes.*

Figure 2: Spreadsheet for Estimating Total Number of Satellite Dishes



Stratum (j)	N_j	n_j	$\boxed{\bar{x}_j}$	$\boxed{s_j^2}$	$\boxed{\left(\frac{N_j}{N}\right)}$	**DxF**	$\boxed{\left(\frac{N_j}{N}\right)^2}$	$\boxed{\frac{s_j^2}{n_j}}$	$\boxed{\frac{N_j - n_j}{N_j}}$	**HxIxJ**
1	400	98	24.1	5,575	0.5722	13.791	0.32747	56.8877551	0.755	14.06473
2	30	10	25.6	4,064	0.0429	1.0987	0.00184	406.4	0.66667	0.499058
3	61	37	267.6	347,556	0.0873	23.353	0.00762	9393.40541	0.39344	28.14555
4	18	6	179	22,789	0.0258	4.6094	0.00066	3798.16667	0.66667	1.679088
5	70	39	293.7	123,578	0.1001	29.412	0.01003	3168.66667	0.44286	14.07285
6	120	21	33.2	9,795	0.1717	5.6996	0.02947	466.428571	0.825	11.34089
Total	**699**				**1**	**77.964**				**69.8022**

$\boxed{\hat{\mu}}$	77.964
$\boxed{\hat{\sigma}_{\hat{\mu}}}$	8.3548
$\boxed{\hat{t} = N\hat{\mu}}$	54,497
$\boxed{\hat{\sigma}_{\hat{t}} = N\,\hat{\sigma}_{\hat{\mu}}}$	5840
$\boxed{\alpha}$	0.05
$\boxed{z_{\alpha/2}}$	1.96
Margin of Err	11,446

We end up with an estimate of 54,497 dishes, with a 95% margin of error equal to 11,446. Again, estimation of totals is not very precise. That does not mean one shouldn't do it, but it's very important to be aware of the margin of error.

Exercises

1. The following table shows Census data for number of people employed, along with sample means from a stratified random sample, where the strata were Provinces or Territories. Give a point estimate of the mean annual income in Canada (in 2001). The answer is a single number.

Province or Territory	Number people employed (Census)	Sample mean reported earnins
Newfoundland	251,545	$24,165
P.E.I.	77,750	$22,303
Nova Scotia	468,825	$26,632
New Brunswick	388,855	$24,971
Quebec	3,815,265	$29,385
Ontario	6,319,530	$35,185
Manitoba	609,575	$27,178
Saskatchewan	534,350	$25,691
Alberta	1,768,435	$32,603
Brit. Columbia	2,128,550	$31,544
Yukon	18,780	$31,526
N.W.T	21,955	$36,645
Nunavut	12,355	$28,215
TOTAL	16,415,785	

2. The United States is divided into four large census regions: Northeast, North Central, South, and West. A random sample of 250 individuals in the labour force were selected from each region. The table below shows percentages unemployed.

Region	Civilian Labour Force (in thousands)	Sample Percent Unemployed
Northeast	27,870.8	4.7%
North Central	34,626.9	5.1
South	53,397.5	4.6
West	34,453.0	4.8

In the calculations that follow, you can omit the finite population correction factors if you wish, because they are so close to one. It won't affect the results.

(a) Give a point estimate of the percent (not proportion) of unemployed persons in the U.S. labour force. The answer is a single number.

(b) Is the sample size big enough to get a margin of error? Answer Yes or No and show your work.

(c) Give the 95% margin of error for the percent unemployed. The answer is a single number.

(d) Give a point estimate of the total number of unemployed persons in the U.S. labour force. The answer is a single number.

(e) Give the 95% margin of error for the total number unemployed. The answer is a single number.

3. Naturally, the U.S. insurance industry is very interested in how long people stay in hospital. They want the stays to be as short (and therefore presumably inexpensive) as possible. For the four large census regions in the US, there are (or were, in 1994) 902 hospitals in the Northeast, 1,704 in the North Central, 2,291 in the South, and 1,126 in the West. A random sample was selected from each region without replacement, and average length of stay was determined from hospital records. The survey was in 1994 too. Results are given in the table below.

```
Analysis Variable : stay Av length of hospital stay, in days

    Region of           N
    country (usa)     Obs            Mean         Std Dev
    -----------------------------------------------------
    Northeast          30      11.0565517       2.6273021

    North Central      50       9.6834375       1.1929378

    South              50       9.1647222       1.2314556

    West               30       8.1137500       1.0031210
    -----------------------------------------------------
```

Give a point estimate and a 95% margin of error for the population mean length of stay for U.S. hospitals in 1994.

3 Quota samples

A quota sample is a stratified but non-probability sample. That is, the population is divided into strata, and then a sample is obtained – somehow – within each stratum. Lots of quota sample come from *panel studies*, in which a large group of consumers are recruited to fill out surveys on an on-going basis, usually in exchange for very modest payment in the form of coupons or free samples. The panels are usually described as "nationally representative," meaning the proportions in various age, sex and maybe income groups is close to census figures for the U. S. or Canadian population.

In quota samples, whether they come from a panel study or not, sample proportions are usually selected to correspond to population proportions (but now always). Invariably, such samples are described as "representative." *You should be aware that the term "representative sample" is not a technical term from Statistics. It is a marketing term.* Someone is trying to sell you data, probably data from a quota sample. Information from a quota sample may be better than unaided intuition, or it may be worse. There is no sure way to tell. For sure, regardless of their age and sex and even regardless of their education, members of a consumer panel are likely to be more literate than average. After all, they are willing to fill out a large number of questionnaires (in English), so it is probably not too difficult for them. The market for books and similar products is consistently over-estimated by panel studies. On the other hand, the market for anti-psychotic drugs is probably under-estimated. Other topics? It's largely guesswork.

Even so, if management decide to take data from quota samples seriously, it is a good idea to apply the methods from this course, and to accompany all estimates with 95% margins of error. At least that way the decision-makers are reminded that the data yield *estimates*, not absolute truth.

If you have data from a quota sample, and the sample sizes are proportional to the population sizes, it is safe to use methods for a simple random sample, because then the confidence intervals will be a little wider and the tests (when we get to them) will be a bit conservative, but that's okay; no harm is done. If the sample sizes are *not* proportional to the population sizes, treat it as a stratified random sample, using the methods from this chapter.